

Noisy Hypotheses in the Age of Discovery Science

Ery Arias-Castro*

We draw attention to one specific issue raised by Ioannidis [17], that of very many hypotheses being tested in a given field of investigation. To better isolate the problem that arises in this (massive) multiple testing scenario, we consider a utopian setting where the hypotheses are tested with no additional bias. We show that, as the number of hypotheses being tested becomes much larger than the discoveries to be made, it becomes impossible to reliably identify true discoveries. This phenomenon, well-known to statisticians working in the field of multiple testing, puts in jeopardy any naive pursuit in (pure) discovery science.

1. Introduction

More than ten years ago now, Ioannidis [17] declared that “most published research findings are false”. If anything, his claim has been corroborated by the empirically observed lack of replicability in the sciences at large. He provided the following argument to support his claim. Consider an ideal setting where, in a certain field of scientific investigation, a number of hypotheses are tested. All tests are assumed to be calibrated at level α and have power $1 - \beta$ for the underlying effect. Let π_1 denote the proportion of false null hypotheses among all the null hypotheses being tested, that is, the proportion of real discoveries to be made. Ioannidis [17] defines the positive predictive value (PPV) as the probability that a given rejection is correct, which he derives under the present setting and obtains

$$\text{PPV} = \frac{(1 - \beta)\pi_1}{(1 - \beta)\pi_1 + \alpha(1 - \pi_1)}. \quad (1)$$

The level α is typically set following standard practice ($\alpha = 0.05$ still being popular in some fields). The power $1 - \beta$ is increasing in α , in the effect size, and in the sample size. The proportion of false nulls π_1 is a consequence of how the hypotheses to be tested are chosen or formulated, and thus arises ‘organically’ from the community of researchers interested in the field of expertise under consideration. From this formula alone Ioannidis [17] derives three corollaries: the first two are about the fact the PPV is increasing in the power and the third one on the fact that the PPV is increasing in the proportion of false nulls (π_1). He encapsulates the impact of the latter in the following statement:

$$\begin{aligned} &\text{“The greater the number and the lesser the selection of tested relationships} \\ &\text{in a scientific field, the less likely the research findings are to be true.”} \end{aligned} \quad (2)$$

This is very easily seen from the fact that

$$\text{PPV} \geq 1/2 \Rightarrow \frac{\pi_1}{1 - \pi_1} \geq \alpha. \quad (3)$$

To ensure that at least half of the claimed discoveries are correct, we see that α needs to be of order at most π_1 (and this is assuming a power of 1). Thus, if π_1 is (very) small, α needs to be set

*University of California, San Diego — <http://www.math.ucsd.edu/~eariasca/>

(very) small, and this presents a problem as it lowers the power. (Remember that the power of a test is decreasing in the level.)

The parameter π_1 could be called the “probing coefficient”. A large π_1 (close to 1) is indicative of a field where hypotheses are generated with care, as might have been the case up to a few decades ago, where data were much more scarce and much more costly to collect and process, and also where the number of researchers (and perhaps the pressure to publish?) was much more limited. A small π_1 is evidence that hypotheses are generated (and tested) with abandon, which might be caricatural of the various ‘omics’, where massive testing is commonplace.

Ioannidis [17] derives the PPV in two other settings: when some form of bias is present and when multiple labs are pursuing the same question independently in parallel. The corresponding formulae lead to three additional corollaries: two are about the fact that PPV is decreasing in the amount of bias and one about the fact that PPV is decreasing in the number of labs performing the same test in parallel.¹ It is generally recognized that bias is an important factor in statistical analyses, due in part to a general preference for positive results (i.e, rejections).

Our main concern here is in exploring the effect that π_1 has on the PPV and its impact on the foundations of so-called ‘discovery science’. In the process, we place the PPV in the context of multiple testing, where it truly belongs, and discuss the issue of distinguishing false from true null hypotheses in this framework.

In a cartoonish simplification, the discovery science paradigm posits the availability of large quantities of data and the scientist — now sometimes referred to as a ‘data scientist’ — lets the data speak for themselves to find meaningful effects. Our message in this paper is that, taken to an extreme, this paradigm is bound to fail if hypotheses are tested with abandon. Taking a multiple testing perspective and drawing from the corresponding literature, we offer some arguments in support of the following statement:

In a field of pure discovery science where very many hypotheses are tested, there is a risk that false nulls are indistinguishable from true nulls.

We argue our point by showing that, as the number of hypotheses increases, the threshold for significance needs to become ever closer to 0, making rejection of both true and false null hypotheses ever harder. This is rather intuitive. Indeed, assuming the null P-values are uniformly distributed in $[0,1]$ under their respective nulls, and if there are a large number m_0 of true null hypotheses, the k -th largest will be close to k/m_0 . Qualitatively speaking, the smallest P-values from the true null hypotheses are ever smaller as their number increases, and the P-values from the false null hypotheses are ‘competing’ with that. We provide some elements of theory that go beyond this simple argument.

2. A utopian model for pure discovery science

To emphasize our point we consider what many would deem a utopian setting where, in a given field of investigation, a number of studies are performed. We assume that each study is preregistered, conducted exactly as planned, and published without delay. We also suppose that each study is based on the same sample size, that each study results in a single bonafide P-value,² that the sought-after effect size is the same, and that the sample size is large enough to result in sufficient power. Moreover, in line with a pure discovery science perspective, we further assume that all studies are, a priori, equally important. We also postulate a fair distribution of resources. Indeed,

¹In [17] it is assumed that only the first lab reaching significance gets to publish its result.

²If p denotes the P-value, this typically means that $\mathbb{P}(p \leq \alpha) \leq \alpha$ for all $\alpha \in [0,1]$ under its null hypothesis. Here we make the simplifying assumption that p is uniformly distributed in $[0,1]$ under its null hypothesis.

Table 1: The setting we consider versus the real state-of-affairs.

Utopia (our setting)	Reality (state-of-affairs)
Each study is preregistered and conducted as planned	The preregistering of studies is not yet common practice [6, 22]
Each study is published	Publication bias can be important [10]
Each study results in a single test	Multiple testing within a single study is common (and sometimes massive) [9, 23]
Each study results in a bonafide P-value	P-values are often approximate or based on unverifiable assumptions, and can be severely biased by informal explorations of the data (data snooping) or flawed by misuse of statistics [12, 26]
Each study is well-powered	Studies are routinely underpowered [5]
Each study tests a different hypothesis	Multiple labs tackle the same question, possibly in different variants and using different tools
Field-wide multiple testing is performed to identify the most significant studies	Published P-values are rarely considered in a larger context (except, possibly, in systematic reviews, but years later)

we imagine the whole field publishing all the results in a single repository and, at the end of each year (say), performing a meta-analysis of all the results published that year. Since in the present setting the P-values carry equal weight, the field as a whole allocates the ‘follow up’ budget equally to further investigate the studies with the most significant P-values. All this is summarized in Table 1, where we also contrast this setting with the reality of many fields of investigation. Of course, the field as a whole would also allocate part of its budget to the generation of new data, new hypotheses, etc, but this will not be of interest here.

In this stylized setting, the sole issue one needs to contend with is the number of hypothesis tests being performed. We implicitly assume this number is large given that we intend to model a field that relies heavily on a data-driven discovery paradigm. Therefore, what remains to be done is to solve the multiple testing problem posed by the very many P-values that are generated. We assume there are m null hypotheses being tested in the period of interest (we said one year above). We denote by $\mathcal{H}_1, \dots, \mathcal{H}_m$ these hypotheses and p_1, \dots, p_m the resulting P-values. (We leave implicit the test statistics that generated these P-values.) We let m_0 and m_1 denote the number of true and false nulls, respectively, and $\pi_0 = m_0/m$ and $\pi_1 = m_1/m$ their proportions.

3. Two interpretations for the PPV

We start with the PPV and derive it in two ways. The first derivation relies on Bayes’ formula and is essentially the derivation of Ioannidis [17], although here we offer a frequentist perspective. The second derivation puts the PPV in the context of multiple testing, where we believe it belongs.

3.1. The derivation of the PPV from Bayes’ formula Ioannidis [17] derives the formula for the PPV, presented in (1), using Bayes’ formula. We provide some details here not provided in

the original article. We note that Ioannidis takes a sort of Bayesian perspective when deriving the PPV, as he (implicitly) assumes that the status of each hypothesis (true or false) is random. We offer a frequentist perspective instead.

Ioannidis postulates that any P-value not exceeding α is deemed significant (and the corresponding null hypothesis is therefore rejected). Suppose that we pick a P-value at random among the P-values that were found significant. We ask ourselves the question: what is the probability that the corresponding null hypothesis is indeed false?

We assume that each P-value is uniformly distributed in $[0, 1]$ under its null hypothesis, so that

$$\mathbb{P}(p_i \leq \alpha) = \alpha, \quad \text{when } \mathcal{H}_i \text{ is true.} \quad (4)$$

We also assume that all tests have the same power $1 - \beta$, so that

$$\mathbb{P}(p_i \leq \alpha) = 1 - \beta, \quad \text{when } \mathcal{H}_i \text{ is false.} \quad (5)$$

Let I be a uniformly distributed index in $\{1, \dots, m\}$. The question we are asking requires the computation of $\mathbb{P}(\mathcal{H}_I^c \mid p_I \leq \alpha)$ where for each i , \mathcal{H}_i^c indicates that \mathcal{H}_i is false. Using Bayes' formula, we derive

$$\mathbb{P}(\mathcal{H}_I^c \mid p_I \leq \alpha) = \frac{\mathbb{P}(p_I \leq \alpha \mid \mathcal{H}_I^c) \mathbb{P}(\mathcal{H}_I^c)}{\mathbb{P}(p_I \leq \alpha \mid \mathcal{H}_I^c) \mathbb{P}(\mathcal{H}_I^c) + \mathbb{P}(p_I \leq \alpha \mid \mathcal{H}_I) \mathbb{P}(\mathcal{H}_I)}. \quad (6)$$

Since the proportions of true and false nulls are π_0 and π_1 ,

$$\mathbb{P}(\mathcal{H}_I) = \pi_0 = 1 - \pi_1, \quad \mathbb{P}(\mathcal{H}_I^c) = \pi_1. \quad (7)$$

Also, by (4),

$$\mathbb{P}(p_I \leq \alpha \mid \mathcal{H}_I) = \alpha, \quad (8)$$

and by (5),

$$\mathbb{P}(p_I \leq \alpha \mid \mathcal{H}_I^c) = 1 - \beta. \quad (9)$$

Plugging this into the large fraction above results in the expression (1). Thus, we have established the following frequentist interpretation of the PPV.

Proposition 1. *Assume we are performing a number of tests at level α , all having power $1 - \beta$. The PPV is the probability that a randomly chosen hypothesis among those rejected is false.*

We note that the independence of the P-values was not used.

Remark 1. Although the setting may appear to be stylized, a similar result holds much more generally. Indeed, suppose that test i is performed at level α_i and has power $1 - \beta_i$. The reader is invited to verify that Proposition 1 still holds in this setting, with (1) unchanged but with α defined as the average level and $1 - \beta$ defined as the average power, meaning

$$\alpha = \frac{1}{m} \sum_{i=1}^m \alpha_i, \quad 1 - \beta = \frac{1}{m} \sum_{i=1}^m (1 - \beta_i). \quad (10)$$

3.2. The PPV as one minus the marginal false discovery rate (mFDR) We present another interpretation of the PPV that puts it in the larger context of multiple testing. It turns out that the PPV can be interpreted as one minus the marginal false discovery rate (mFDR) [27] of the naive multiple test that rejects every null hypothesis with a P-value below α — see (12) below. We provide some background for the reader not already familiar with multiple testing.

Table 2: Some notation commonly used in multiple testing. All the quantities, except for m, m_0, m_1 are random and depend on the (implicit) multiple testing procedure in use. In this notation, $\pi_0 = m_0/m$ is the proportion of true nulls and $\pi_1 = m_1/m$ is the proportion of false nulls.

	True Null	False Null	Total
Rejection	V	S	R
No Rejection	U	T	$m - R$
Total	m_0	m_1	m

In the particular setting considered here, where the P-values are assumed independent, we may without loss of generality define a multiple test as a function of the P-values p_1, \dots, p_m that returns an index $\mathcal{R} \subset \{1, \dots, m\}$ indicating the hypotheses to be rejected. In formula, the procedure rejects \mathcal{H}_i if $i \in \mathcal{R}$. Applying a given multiple test results in the situation described in Table 2.

A single test is evaluated based on its probabilities of type I and type II errors, or equivalently, its size and power. A good test is a test with small size and large power. Analogous measures need to be considered when evaluating a multiple test: one based on the number of rejected true nulls and another one based on the number of false nulls rejected (V and S in Table 2). A good multiple test is a multiple test with small number of rejected true nulls and large number of rejected false nulls. The mFDR provides a notion of size for multiple tests. Given a multiple test (left implicit) resulting in the situation of Table 2, it is defined as

$$\text{mFDR} = \frac{\mathbb{E}(V)}{\mathbb{E}(R)}. \quad (11)$$

Ioannidis [17] implicitly considers the α -threshold multiple test, defined by

$$\mathcal{R}_\alpha = \{i : p_i \leq \alpha\}, \quad (12)$$

where α is given.

Proposition 2. *Assume that each test has power $1 - \beta$ at level α . Then*

$$\text{PPV} = 1 - \text{mFDR}, \quad (13)$$

where mFDR is the mFDR for the α -threshold multiple test (12).

The proof is straightforward. Indeed, the expected number of rejections is

$$\mathbb{E}(R) = m_0\alpha + m_1(1 - \beta), \quad (14)$$

and the expected number of incorrect rejections is

$$\mathbb{E}(V) = m_0\alpha. \quad (15)$$

Thus

$$\text{mFDR} = \frac{m_0\alpha}{m_0\alpha + m_1(1 - \beta)} = 1 - \frac{\pi_1(1 - \beta)}{\pi_0\alpha + \pi_1(1 - \beta)}, \quad (16)$$

after dividing by m in the numerator and denominator, and we conclude with the fact that $\pi_0 = 1 - \pi_1$.

4. Optimal multiple testing procedures

The last section gives a multiple testing interpretation of the PPV as one minus the mFDR that results from using the α -threshold multiple test (12) in a context where the tests have power $1 - \beta$ at level α . We already discussed in (3) some implications for using this multiple test, but now that we placed the problem in the larger context of multiple testing, one has to wonder if a more sophisticated approach would improve the situation. Indeed, in our utopian setting, the field as a whole has all interest in using a good multiple test to decide, in a principled way, which studies are significant and worth of further investigation.

Surely, there is a large and growing statistical literature on the topic of multiple testing, offering a number of methods. We refer the reader to some books [7, 9] and some papers surveying the field [11, 24].

4.1. The family-wise error rate (FWER) The family-wise error rate (FWER) of a given multiple test is defined as the probability of an incorrect rejection, or in the notation of Table 2,

$$\text{FWER} = \mathbb{P}(V \geq 1). \quad (17)$$

The α -threshold multiple test is still popular, for example, in genetics. It is discussed in [2] in the context of genome-wide association studies (GWAS), where α is said to be typically set to 5×10^{-8} . The rationale is that a typical study probes on the order of 10^6 (one million) SNP's, so that the threshold is set by Bonferroni's method ($0.05/10^6 = 5 \times 10^{-8}$). This guaranties control of the FWER in a given study. However, it does not control for the fact that hundreds, if not thousands of such studies are performed each year.

4.2. The false discovery rate (FDR) For many decades the FWER was the main measure of size for multiple tests and apparently remains quite popular. However, in contexts where a large number of tests are performed (large m), controlling the FWER is known to severely limit the (multiple) power. This lead Benjamini and Hochberg [3] to introduce the false discovery rate (FDR), which in the notation of Table 2 is defined as

$$\text{FDR} = \mathbb{E} \left(\frac{V}{\max(R, 1)} \right). \quad (18)$$

This is now a popular notion of size for multiple tests, for example, in genetics [4, 9].

The mFDR is a later variant of the FDR [27]. Although the mFDR is conceptually simpler than the FDR, the former cannot be properly controlled when all the null hypotheses are true ($m_0 = m$) while it is still possible to control the latter. That said, they are known to be equivalent under somewhat mild assumptions when the number of tests is large. Nevertheless, the FDR appears more popular in practice and we consider control of the FDR in what follows.

4.3. The Benjamini-Hochberg (BH) procedure in the normal means model Table 3 describes the famous procedure of Benjamini and Hochberg [3], known to control the FDR (at a prescribed γ) when the P-values are independent.³ Note that this multiple test is of threshold type, meaning of the form

$$\mathcal{R} = \{i : p_i \leq \tau\}. \quad (19)$$

³The method is also known to control the FDR at the prescribed level under some forms of positive dependence.

Table 3: The Benjamini and Hochberg [3] (BH) multiple test.

p_1, \dots, p_m denote the P-values; γ denotes the desired FDR level
1: Order the P-values, obtaining $p_{(1)} \leq \dots \leq p_{(m)}$.
2: Let \hat{s} denote the largest s such that $p_{(s)} \leq s\gamma/m$.
3: Reject the nulls corresponding to the \hat{s} smallest P-values.

For the BH multiple test, the threshold is

$$\tau_\gamma = \hat{s}\gamma/m, \quad \text{where } \hat{s} \text{ is defined in Table 3.} \quad (20)$$

Note that it is a function of the P-values.

The normal means model assumes that each test being performed is a Z-test and that any underlying effect only affects the mean. In detail, we assume that for each $i \in \{1, \dots, m\}$, we have available a Z-ratio X_i and that we reject for large values of X_i . The P-value is computed assuming that X_i is standard normal under \mathcal{H}_i , meaning $p_i = 1 - \Phi(X_i)$, where Φ denotes the standard normal distribution function. We further assume that the model is indeed accurate in that $X_i \sim \mathcal{N}(0, 1)$ under \mathcal{H}_i , while $X_i \sim \mathcal{N}(\mu, 1)$ under \mathcal{H}_i^c , where $\mu > 0$ (since we are rejecting for large values of X_i). For example, Ingster [16] and Donoho and Jin [8] considered the problem of testing the global null hypothesis $(\cap_{i=1}^m \mathcal{H}_i)$ in this setting. Note that in this model $p_i \sim \text{Unif}(0, 1)$ under \mathcal{H}_i , while $p_i \sim F_\mu$ under \mathcal{H}_i^c , where $F_\mu(u) := 1 - \Phi(\Phi^{-1}(1 - u) - \mu)$.

In an asymptotic setting where $m \rightarrow \infty$ but everything else remains essentially constant, we have the following consequence of a more general result of Genovese and Wasserman [13]. We first remark that for any $\mu > 0$ and any $\lambda > 1$, there is a unique point $u \in (0, 1)$ such that $F_\mu(u) = \lambda u$, which we denote by $u(\mu, \lambda)$. We note that $u(\mu, \lambda)$ is a decreasing function of λ and that $u(\mu, \infty) = 0$.

Proposition 3. *Consider an asymptotic setting where $m \rightarrow \infty$ while $\pi_1 = \pi_1(m) \rightarrow \pi_1^\infty \in (0, 1)$ and $\mu = \mu(m) \rightarrow \mu^\infty > 0$. Then τ_γ defined in (20) converges in probability to $u(\mu^\infty, \lambda^\infty)$ where $\lambda^\infty = 1 + (1 - \gamma)/\gamma\pi_1^\infty$.*

This result implies that, as the proportion of false nulls π_1 becomes negligible ($\pi_1 \rightarrow 0$), the threshold for the BH multiple test converges to zero ($\tau_\gamma \rightarrow 0$). The conclusion is thus morally the same as with the α -threshold multiple test: a small π_1 makes rejection more difficult, that is, if we use the BH procedure for controlling the FDR.

4.4. Optimality in the normal means model It is possible to study questions of optimality in the context of multiple testing in the normal means model. For example, Jin and Ke [18] do this in a more general minimax framework⁴ that includes the present context. We present here an alternative, and arguably simpler approach, that based on oracle information [1, 19, 28].

Suppose that an oracle tells us exactly which null hypotheses are false. With that information, we would just reject precisely these hypotheses. But knowing this does not enlighten the discussion! Imagine, however, that we are restricted to use a threshold procedure, meaning that the rejection set needs to be of the form (19). This means that if we reject \mathcal{H}_i then we must reject \mathcal{H}_j for all j such that $p_j \leq p_i$. Remember that in our setting there is nothing that distinguishes the null hypotheses a priori, so doing this is rather natural. The α -threshold multiple test (implicitly considered by Ioannidis) and the Benjamini-Hochberg procedure are certainly of this form. Now, what is the

⁴The setting in that paper is that of classification, which is intimately related to the problem of multiple testing.

best that we can do with this oracle information (i.e., full disclosure) under the constraint that we must use a threshold procedure? A good candidate is the largest threshold that controls the false discovery proportion (FDP) at the desired level. For a given multiple test, in the notation of Table 2,

$$\text{FDP} = \frac{V}{\max(R, 1)}, \quad (21)$$

so that $\text{FDR} = \mathbb{E}(\text{FDP})$. Let V_t and R_t , and FDP_t , denote the corresponding quantities when we select the P-values not exceeding t . Define the oracle threshold procedure as

$$\tau_\gamma^* = \max \{t : \text{FDP}_t \leq \gamma\}. \quad (22)$$

(We note that the procedure we consider in [1] is slightly different.)

Following the same arguments leading to Proposition 3, we can derive the asymptotic behavior of the oracle threshold multiple test.

Proposition 4. *In the same setting as Proposition 3, the oracle threshold τ_γ^* also converges in probability to the same limit $u(\mu^\infty, \lambda^\infty)$.*

Thus the BH procedure achieves the same first-order asymptotic performance as the oracle procedure. This is true in a setting where the model parameters remain constant as the number of hypotheses increases. In [1] we consider the situation where the model parameters are allowed to change with m and arrive at essentially the same conclusion.⁵

If the oracle procedure is the best we can hope to achieve, then we can conclude, as before, that in order to control the FDR the threshold for rejection needs to be taken closer and closer to 0 as the number of hypotheses increases. Because of that, there is a real danger that the false nulls drown in the midst of a large number of true null hypotheses.

5. Discussion

By properly placing the PPV in the context of multiple testing, we provided some additional theoretical backing to Ioannidis' claim (2). Perhaps it is no longer just about what each lab does in its corner; perhaps it also matters what other labs are doing. For if some labs are repeatedly probing the world more or less at random, without much thought, the hypotheses they test and the P-values they publish will act as noise and possibly overwhelm the field to a point where it becomes impossible, at least with statistical methodology alone, to separate the false nulls from the true nulls.

Field-wide control of false discovery rate? At a higher level, we question whether research can continue to be done independently by a large number of researchers without some sort of central decision making on what constitutes important questions. There are some initiatives in this direction [22]. Rosset et al. [25], very recently, proposed a control via access to a large database. Note that this proposal requires the collection of new samples so that the database is continually enriched and FDR control remains possible.

Blame it on the P-values? With the bad reputation that P-values are getting and the long-standing controversy around them [20, 21, 29, 30], one may wonder if all this can also be blamed on them. For example, one may wonder if the use of Bayes factors [14, 15], or even confidence intervals, would solve the problem. The answer is a resounding *no* if, as we consider here, the goal is to identify the false nulls with some degree of accuracy.

⁵The BH procedure is likely optimal also in a minimax sense, although this has not been proved.

Acknowledgments I am grateful to my colleagues, Jeff Rabin and Ian Abramson, for extensive discussions on the topic and for reading a draft of the manuscript. This work was partially supported by a grant from the US National Science Foundation (DMS 1223137).

References

- [1] E. Arias-Castro and S. Chen. Distribution-free multiple testing. *arXiv preprint arXiv:1604.07520*, 2016.
- [2] G. S. Barsh, G. P. Copenhaver, G. Gibson, and S. M. Williams. Guidelines for genome-wide association studies. *PLoS Genetics*, 8(7):e1002812, 2012.
- [3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 289–300, 1995.
- [4] W. S. Bush and J. H. Moore. Genome-wide association studies. *PLoS Computational Biology*, 8:2822, 2012.
- [5] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.
- [6] K. Dickersin and D. Rennie. Registering clinical trials. *JAMA*, 290(4):516–523, 2003.
- [7] T. Dickhaus. *Simultaneous Statistical Inference: With Applications in the Life Sciences*. Springer Science & Business Media, 2014.
- [8] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.
- [9] S. Dudoit and M. J. Van Der Laan. *Multiple testing procedures with applications to genomics*. Springer Science & Business Media, 2008.
- [10] K. Dwan, D. G. Altman, J. A. Arnaiz, J. Bloom, A.-W. Chan, E. Cronin, E. Decullier, P. J. Easterbrook, E. Von Elm, and C. Gamble. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One*, 3(8):e3081, 2008.
- [11] H. Finner, K. Strassburger, et al. The partitioning principle: a powerful tool in multiple decision theory. *The Annals of Statistics*, 30(4):1194–1213, 2002.
- [12] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Technical report, Columbia University, 2013.
- [13] C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.
- [14] S. N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999.
- [15] S. N. Goodman. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12):1005–1013, 1999.
- [16] Y. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.
- [17] J. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.
- [18] J. Jin and T. Ke. Rare and weak effects in large-scale inference: methods and phase diagrams. *arXiv preprint arXiv:1410.4578*, 2014.
- [19] N. Meinshausen, M. H. Maathuis, P. Bühlmann, et al. Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391, 2011.
- [20] P. A. Murtaugh. In defense of p values. *Ecology*, 95(3):611–617, 2014.
- [21] R. Nuzzo. Statistical errors. *Nature*, 506(7487):150–152, 2014.
- [22] Open Science Collaboration. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6):657–660, 2012.
- [23] T. K. Rice, N. J. Schork, and D. Rao. Methods for handling multiple testing. *Advances in Genetics*, 60:293–308, 2008.

- [24] E. Roquain. Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique*, 152(2):3–38, 2011.
- [25] S. Rosset, E. Aharoni, and H. Neuvirth. Novel statistical tools for management of public databases facilitate community-wide replicability and control of false discovery. *Genetic Epidemiology*, 38(5): 477–481, 2014.
- [26] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11): 1359–1366, 2011.
- [27] J. D. Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, 2007.
- [28] W. Sun and T. T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.
- [29] D. Trafimow and M. Marks. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2, 2015.
- [30] R. L. Wasserstein and N. A. Lazar. The ASA’s statement on p-values: context, process, and purpose. *The American Statistician*, 2016. In press.